DMPA Lab Project: Capturing Every Emotion

The Project was equally Contributed By:

41) Niraj Yagnik - 170953170

42) Manthan Jain - 170953172

59) Tulika Banerjee - 170953246

DMPA Project(2020)

Results and Analysis

<u>Capturing Every Emotion</u>: *An In-depth Facial Expression Study Using Multimodal Sentiment Analysis and Facial Emotion Recognition of Real-Time Video Using Neural Networks*

Brief Overview

The project attempts to build a robust and efficient analytical study on human facial behavior by not only analyzing how the human face functions during a conversation but also research ways to find a correlation between what people say and the facial expressions they make when saying it. The purpose of this research is to help boost the accuracy of sentiment prediction by also analyzing the facial features along with the person's speech transcriptions.

Results:

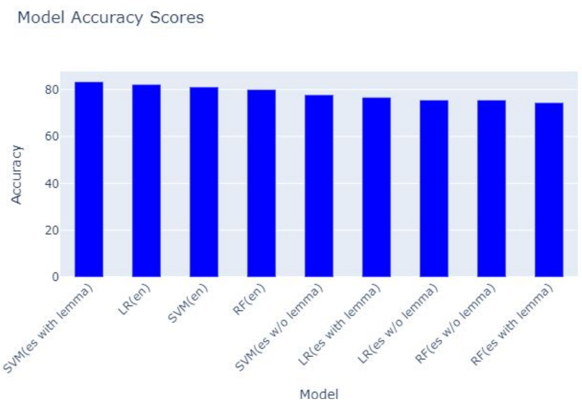
1. Part 1*: Multimodal Sentiment Analysis(Video + Text Data)*

Bried Details of the Implementation:

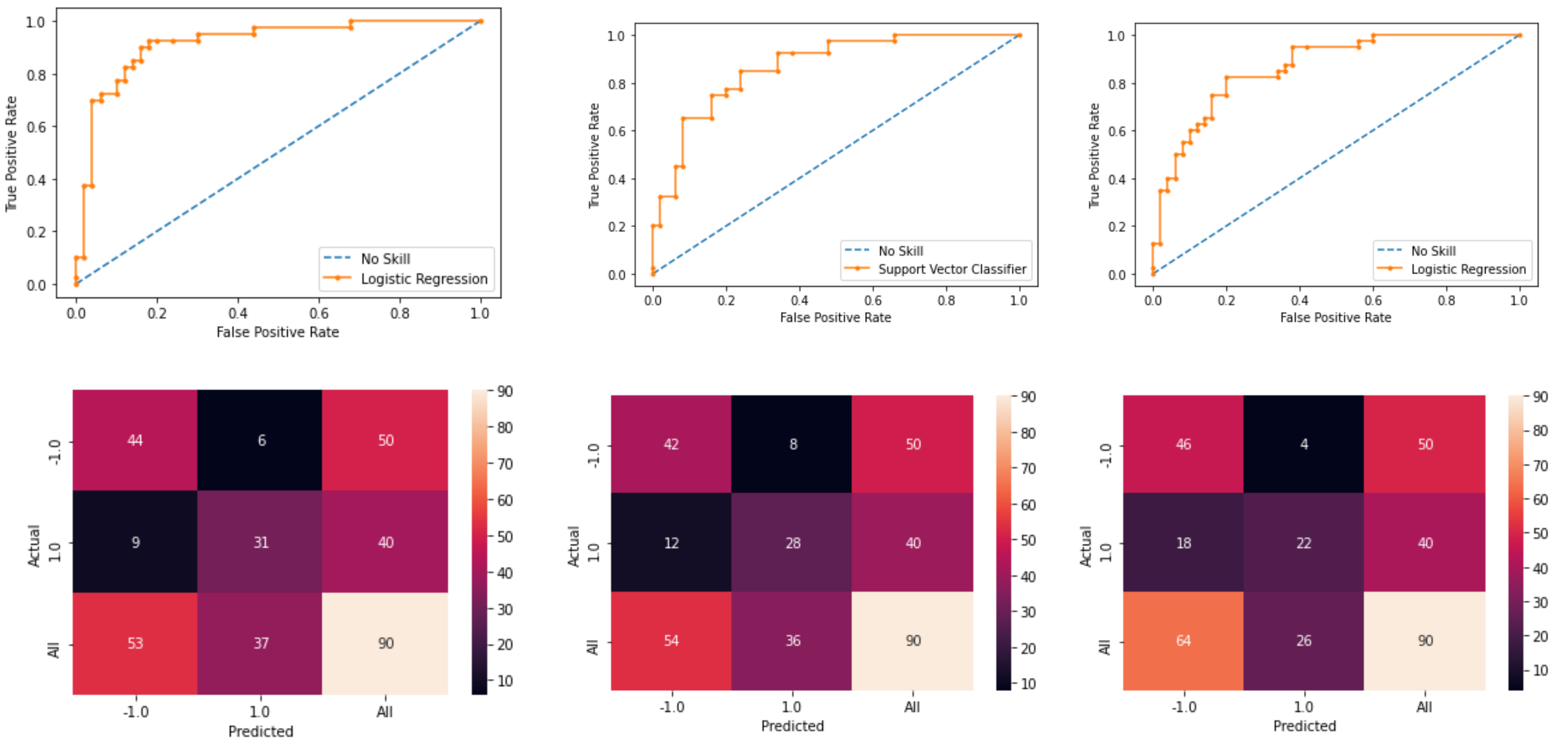
- We use MOUD Dataset for our experimentation which is a collection of video reviews, segmented at utterance level, transcribed, and annotated for the sentiment.
- We leverage the power of Machine Learning Algorithms to prove that the addition of facial features improves the results of the initial sentiment analysis using just text data.
- We also conduct a deep study on a corollary where textual features of language B and non-verbal cue features of language A are analyzed together in an attempt to find the variation in the results which can then be extrapolated to the initial situation, thereby providing useful insight into the impact such a cultural shift in non-verbal cues can have on multimodal analysis.

Text Data Sentiment Analysis

Top Scores Achieved(Over 20 models were run)



ROC Curves and Confusion Matrix for some of the models

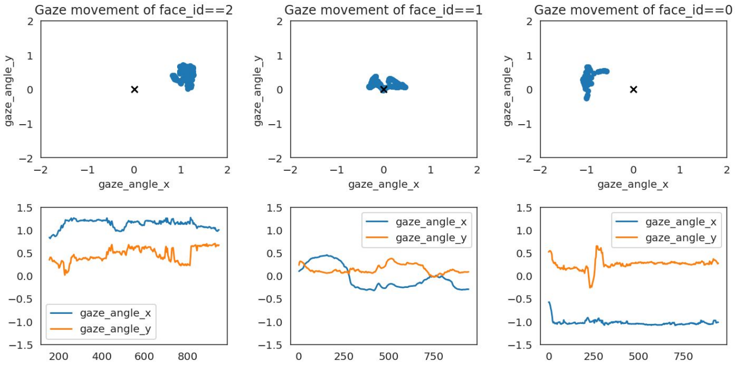


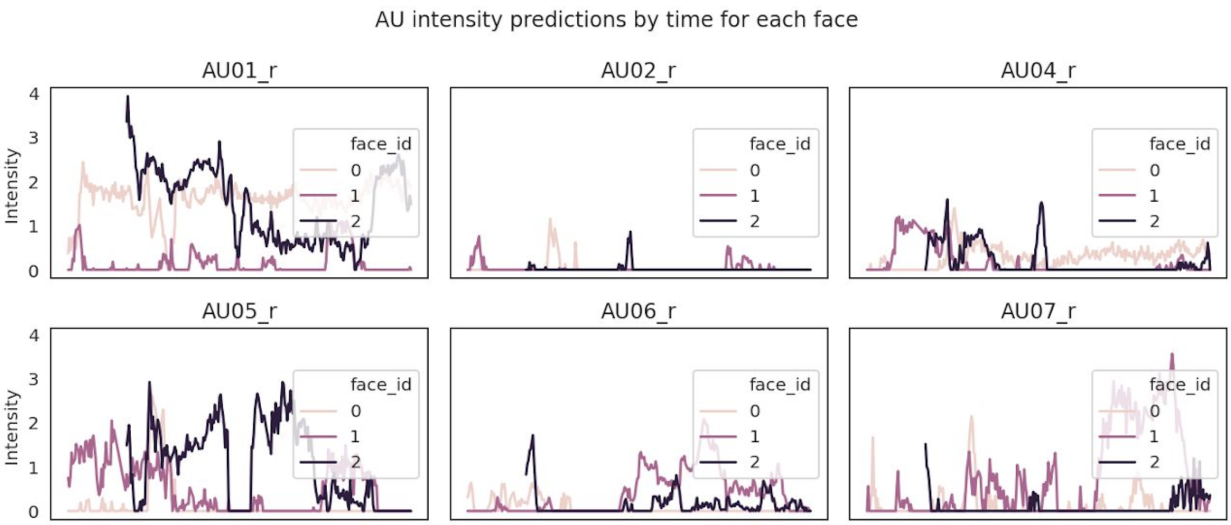
LR(en) SVM(es w/o lemma) LR(es)

- Over 40 Machine Learning models were trained and tested to achieve the best possible results. Algorithms were trained on both Spanish and English text data to gain an insight into the changes in sentiments that could occur due to the translation of the text.
- SVM with Spanish Text undergoing Lemmatization gives the highest scores but in general, the models with English texts perform better than Spanish Text.

Video Data Sentimental Analysis

- In this section, we augment the power of OpenFace API, a state-of-the-art facial feature extraction algorithm. The features extracted for each utterance are averaged over all the valid frames and each row with the averaged value over the period of utterance is associated with sentiment associated with that particular utterance.
- Graphs for the features extracted by the OpenFace API





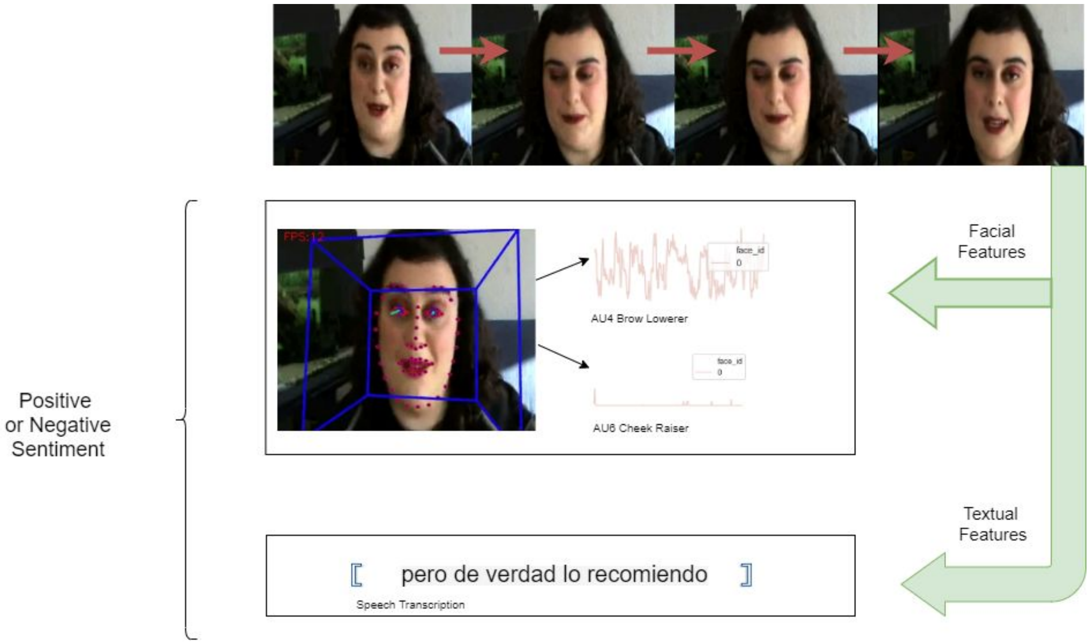
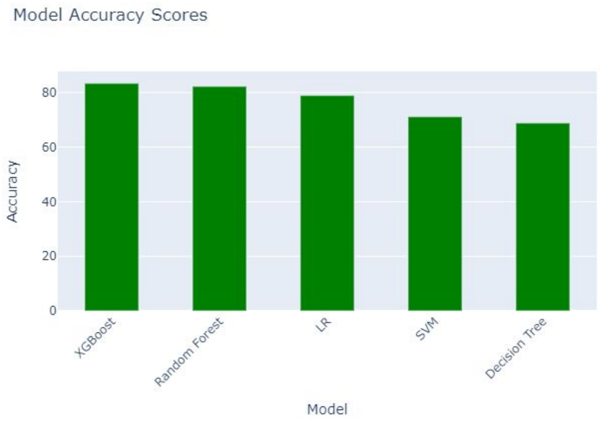


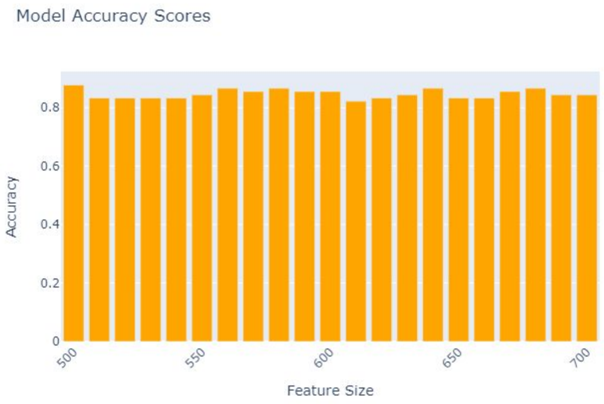
Figure illustrating feature extraction performed by OpenFace

- Model Score



Video+Text Data Sentimental Analysis

- The text dataframe and video dataframe are horizontally stacked at point of overlapping Utterances. Implementation: Over 20 models were implemented to check for the efficacy of the implementation. It was performed for both Spanish and English Text.
- Feature Selection Performed Using Recursive Feature Elimination.

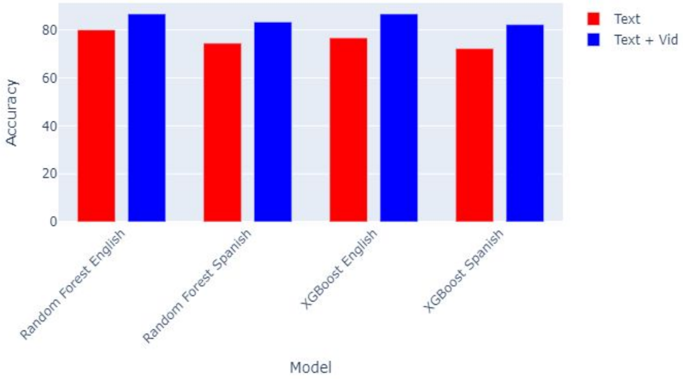


Model Accuracy Scores different Feature Size

Results illustrating the improvement in accuracy scores when adding Video Data

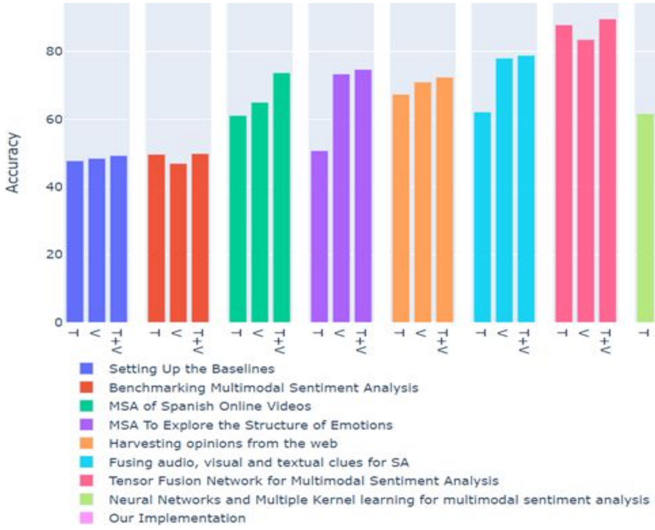
Spanish v/s English Models Performance

	Text	Video	Text+Video
Random Forest (English)	0.744	0.833	0.833
Random Forest (Spanish)	0.8	0.833	0.866
XGBoost (English)	0.766	0.822	0.822
XGBoost (Spanish)	0.722	0.822	0.866



Comparing our model scores with scores of Research Paper Implementations

Model/Paper Name	**T**	**V**	**T+V**
Benchmarking Multimodal Sentiment Analysis	49.6	46.9	49.8
Multimodal Sentiment Analysis of Spanish Online Videos	61.04	64.94	73.68
Multimodal Sentiment Analysis To Explore the Structure of Emotions	50.66	73.33	74.66
Towards multimodal sentiment analysis: harvesting opinions from the web	67.31	70.94	72.39
Fusing audio, visual and textual clues for sentiment analysis	62.11	78.00	78.80
Tensor Fusion Network for Multimodal Sentiment Analysis	87.83	83.48	89.57
Neural Networks and Multiple Kernel Learning for multimodal sentiment analysis	61.60	76.48	77.17
Our Implementation	**83.33**	**84.44**	**88.89**



Analysis and Conclusions

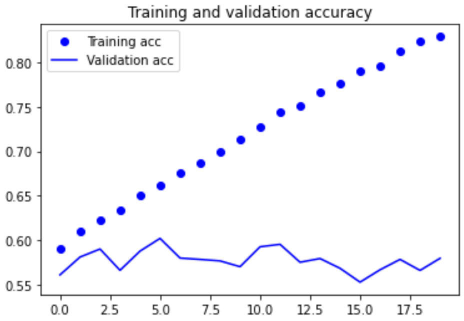
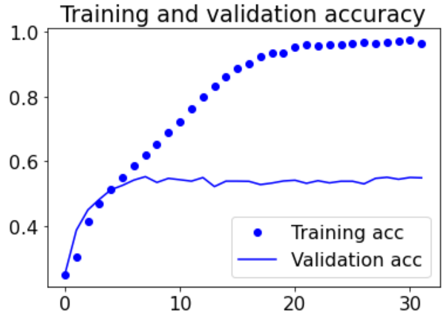
- When analyzed using text alone, the English translated text, in general, outperforms the Spanish models.
- However, when running along with the video, the Spanish prediction models immediately shoot up in accuracy, whereas the English models perform poorly relative to the Spanish Model As we can see, Spanish is almost always less accurate than English, but taking best models for each, we see it is comparable.
- However, as we can see once we add the Visual features, the disparity between the models' accuracy grows manifold.
- This goes to show how the facial feature cues that accompany the original Spanish speech are in sync with what is being said, whereas the English speech and the cues do not match up, although the English speech relays the exact same sentiment as its Spanish counterpart.

2. Part 2: Real-time Emotion Recognition

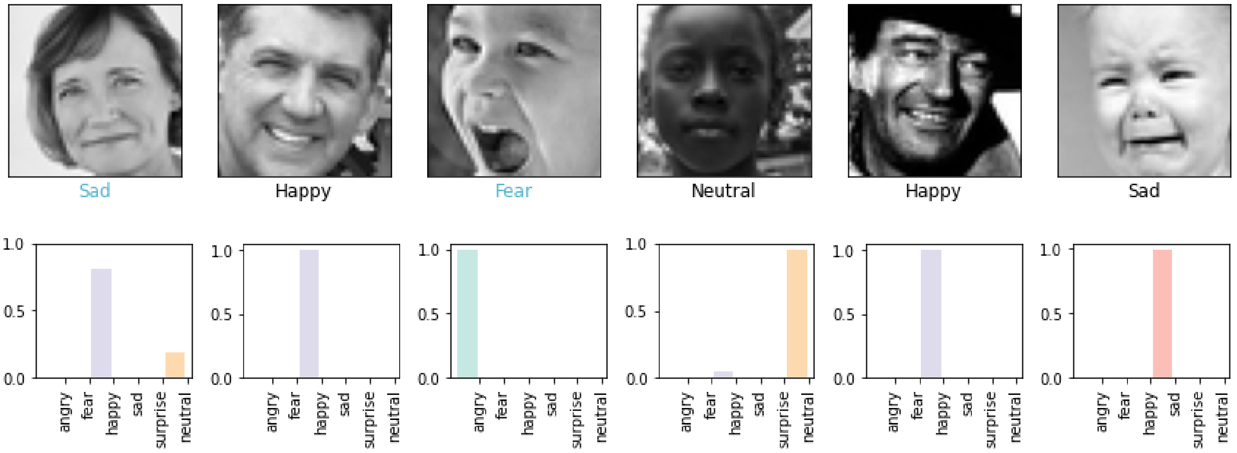
Brief Overview:

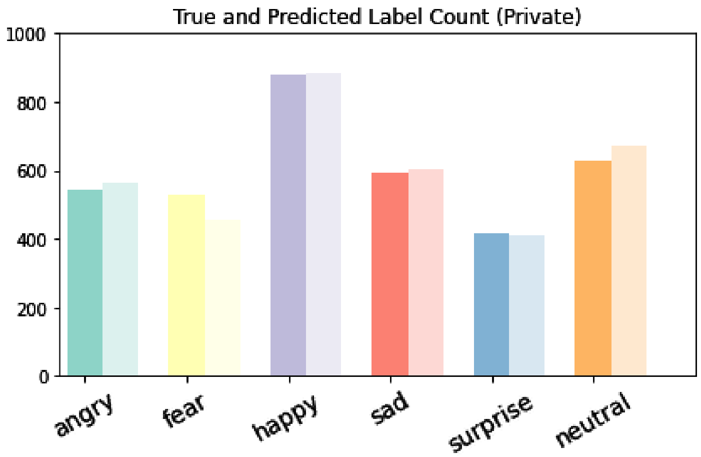
- A face emotion recognition system comprises a two-step process i.e. face detection (bounded face) in image followed by emotion detection on the detected bounded face.
- A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other.
- We leverage the power of CNN architecture for our experimentations.
- *Dataset*: The training set consists of 35,888 examples. train.csv contains two columns, "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image(the image(Anger, Suprised, Disgust, Happy, Sad, Neutral, Scared)

Results:

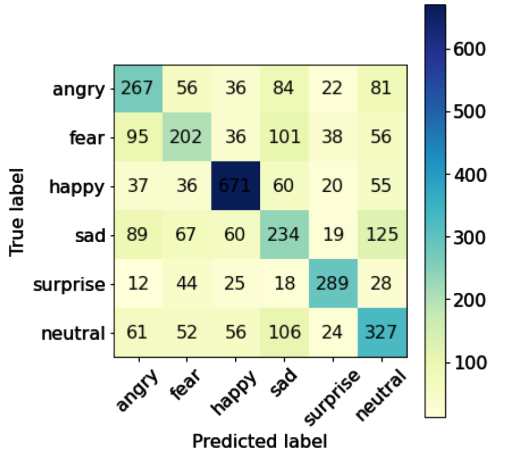
Model Architecture:Xception CNN Model	*Model Architecture: Fully Connected CNN Architecture*
 Accuracy Score: : 57.94	 Accuracy Score: : 55.45

Testing Results:





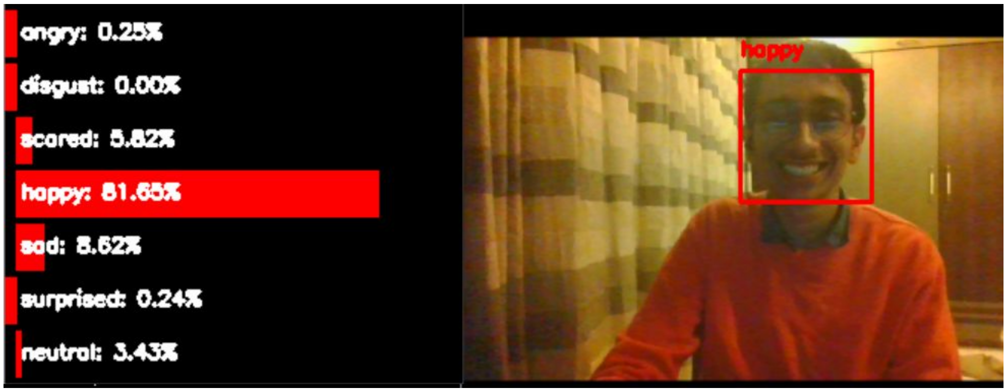
Bar Graph indicating True v/s predicted label count

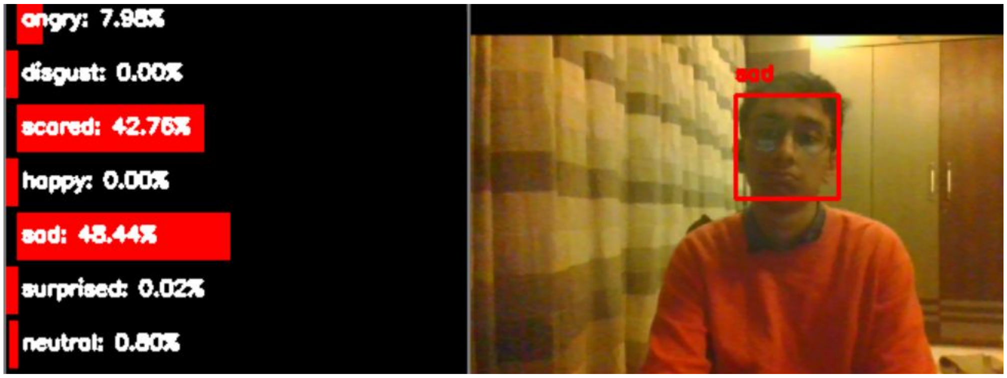


Confusion Matrix

Real-Time Analysis

- Haar feature-based cascade classifier is used for face detection. The CNN model trained is saved and used for detecting the emotion of the person in real-time using video handling functions in OpenCV.


angry: 0.25%
disgust: 0.00%
scared: 5.82%
happy: 81.65%
sad: 8.62%
surprised: 0.24%
neutral: 3.43%
happy


angry: 7.98%
disgust: 0.00%
scared: 42.76%
happy: 0.00%
sad: 48.44%
surprised: 0.02%
neutral: 0.80%
sad